The Effect of Pixel-Level Fusion on Object Tracking in Multi-Sensor Surveillance Video

N. Cvejic, S. G. Nikolov, H. D. Knowles, A. Łoza, A. Achim, D. R. Bull and C. N. Canagarajah
Centre for Communications Research, University of Bristol
Merchant Venturers Building, Woodland Road, Bristol BS8 1UB, United Kingdom
n.cvejic@bristol.ac.uk

Abstract

This paper investigates the impact of pixel-level fusion of videos from visible (VIZ) and infrared (IR) surveillance cameras on object tracking performance, as compared to tracking in single modality videos. Tracking has been accomplished by means of a particle filter which fuses a colour cue and the structural similarity measure (SSIM). The highest tracking accuracy has been obtained in IR sequences, whereas the VIZ video showed the worst tracking performance due to higher levels of clutter. However, metrics for fusion assessment clearly point towards the supremacy of the multiresolutional methods, especially Dual Tree-Complex Wavelet Transform method. Thus, a new, tracking-oriented metric is needed that is able to accurately assess how fusion affects the performance of the tracker.

1. Introduction

Multi-sensor data often presents complementary information about the region surveyed and data fusion provides an effective method to enable comparison, interpretation and analysis of such data. The aim of video fusion, apart from reducing the amount of data, is to create new videos that are more suitable for human perception, and for further video processing tasks such as segmentation, object detection or target recognition.

The fusion of multi-modal video sources is becoming increasingly important for surveillance purposes, navigation and object tracking applications. For example, combining visible and infrared sensors produces a fused image constructed from a combination of features, which enables improved detection and unambiguous localisation of a target (represented in the IR image) with respect to its background (represented in the visible image). A human operator using a suitably fused representation of visible and IR images may therefore be able to construct a more complete and accurate mental representation of the perceived scene, resulting in a larger degree of situation awareness [16].

Recently there has been an increased interest in object tracking in video sequences supplied by either a single camera or a network of cameras [7, 1, 3, 12, 11, 15]. Reliable tracking methods are of crucial importance in many surveillance systems as they enable human operators to remotely monitor activity across areas of interest and help the surveillance analyst with the decision-making process.

There are three levels of interaction between video fusion and tracking algorithms. At the first level, fusion is performed at the pixel level and tracking is performed using the fused video sequence. At the second level, one or more features/cues are extracted from the input videos and tracking is implemented using the fused cues. Finally, at the third level tracking can be performed on all input videos, using different cues, and fusion of tracking trajectories is performed at the decision level. The interaction between fusion and tracking has been explored extensively at the second and third levels, whereas the impact of pixel-level fusion on the performance of trackers has been studied so far in only a few publications (see for instance [10]).

In this paper, we investigate how pixel-level fusion of videos from VIZ and IR cameras influences the process of object tracking in surveillance video. We compare the tracking results from the fused sequences with the results obtained from separate IR and VIZ videos. In addition, we evaluate the performance of the state-of-the-art video fusion methods by computing standard fusion metrics on a frame-by-frame basis.

2. Object Tracking Using Particle Filtering

Particle Filtering (PF) was first used for tracking through a video sequence by Isard et al. [7]. The advantage of particle filtering is that the restrictions of linearity and Gaussianity imposed by the Kalman filter are relaxed [1]. This is a most useful property in the field of video tracking, where there can be significant clutter which results in highly non-

1-4244-1180-7/07/$25.00 ©2007 IEEE

Gaussian likelihood densities. Whereas the work of Isard et al. [7] focused on tracking contours, more recent work has been on features or properties of the target. Comaniciu et al. [3] proposed using the Bhattacharyya distance between two colour histograms to determine the likelihood of a given location. The Mean Shift algorithm was used to find the most likely target location. The idea of using the Bhattacharyya distance within the particle filtering framework was first presented by Pérez et al. [12] and Nummiaro et al. [11].

Shen et al. [15] fuse colour and a simplified shape contour within a particle filtering framework. The cues are considered independent, and the weights for each cue are adapted depending on how well each cue agrees with the aggregate result. Brasnett et al. [2] extend this approach to use colour, edges and texture. The weightings for the cues come from the current frame rather than the previous frame (as in [15]), and are determined by the minimum distance between the target histograms and the observed histograms.

A different type of cue was proposed by Łoza et al. [9], where the structural similarity measure proposed by Wang et al. [17] is used. This combines measures of similarity for the mean / luminance, contrast, and structure or correlation.

Within the context of recursive Bayesian filtering, we are trying to estimate the probability of the state at time k ($\mathbf{x}_k$), given the current measurement ($\mathbf{z}_k$):

$$p(\mathbf{x}_k|\mathbf{z}_k) = \frac{p(\mathbf{z}_k|\mathbf{x}_k)p(\mathbf{x}_k|\mathbf{z}_{k-1})}{p(\mathbf{z}_k|\mathbf{z}_{k-1})} \tag{1}$$

A particle filter aims to represent $p(\mathbf{x}_k|\mathbf{z}_k)$ with a set of particles. Each particle has an associated weight, w_k^i where i is the index, and N_s is the number of samples or particles. Using the standard Sampling-Importance-Resampling (SIR) filter [1], the weights are given by:

$$w_k^i = p(\mathbf{z}_k|\mathbf{x}_k^i) \tag{2}$$

and thus the posterior is approximated by:

$$p(\mathbf{x}_k|\mathbf{z}_k) \approx \sum_{i=1}^{N_s} w_k^i \delta(\mathbf{x}_k - \mathbf{x}_k^i) \tag{3}$$

2.1. Colour Cue

When using colour histograms for tracking, the question remains as to how to use the histogram to compute $p(\mathbf{z}_k|\mathbf{x}_k)$. The following method was first proposed by Pérez et al. [12]. If $\mathbf{h}_{\text{target}}$ is the reference or target histogram i.e. our model of the target's appearance, and $\mathbf{h}_{\text{prop}}$ is the proposal histogram, then the Bhattacharyya distance is given by:

$$\rho(\mathbf{h}_{\text{target}}, \mathbf{h}_{\text{prop}}) = \sum \sqrt{h_{\text{target}} \times h_{\text{prop}}} \tag{4}$$

This can then be converted into a metric [3]:

$$d_{\text{colour}}(\mathbf{h}_{\text{target}}, \mathbf{h}_{\text{prop}}) = \sqrt{1 - \rho((\mathbf{h}_{\text{target}}, \mathbf{h}_{\text{prop}}))} \tag{5}$$

The likelihood function for the colour cue is then defined as [12]:

$$\mathcal{L}_{\text{colour}}(\mathbf{z}|\mathbf{x}) \propto \exp\left[-\frac{d^2_{\text{colour}}(\mathbf{h}_{\text{target}}, \mathbf{h}_{\text{prop}})}{2\sigma^2}\right] \tag{6}$$

where σ is chosen empirically.

2.2. SSIM Cue

The structural similarity measure was first proposed by Wang et al. [17]; it was first used for tracking in [9]. Its purpose is to provide a measure of how similar two image patches X and Y are to each other. The three criteria for measuring the similarity are luminance, contrast, and structure (or correlation):

$$Q(X,Y) = \frac{2\mu_x\mu_y}{\mu_x^2 + \mu_y^2} \frac{2\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2} \frac{\sigma_{xy}}{\sigma_x\sigma_y} \tag{7}$$

The range of values for Q are $Q \in [-1 \ldots 1]$, where $Q = 1$ means the that two patches are identical. Because there is no clear interpretation as to how to treat values of Q that are less than zero, we set any of these values to zero. In a manner similar to (6), we calculate the likelihood as:

$$\mathcal{L}_{\text{SSIM}}(\mathbf{z}|\mathbf{x}) \propto \exp\left[-\frac{(1 - Q(X,Y))^2}{2\sigma^2}\right] \tag{8}$$

2.3. Cue Fusion

It is assumed that the two cues described in Sections 2.1 and 2.2 are independent. The overall likelihood is calculated as:

$$\mathcal{L}(\mathbf{z}|\mathbf{x}) = \mathcal{L}_{\text{SSIM}}(\mathbf{z}|\mathbf{x}) \times \mathcal{L}_{\text{colour}}(\mathbf{z}|\mathbf{x}) \tag{9}$$

While this assumption may not be true in all cases, it was found that the fused tracker typically gives an improved performance over the single cue case, and consequently this approach was used throughout the rest of the paper.

3. Video Fusion

The videos described in Section 4.1 were fused using the following methods: Averaging Technique (AVE), Laplacian Pyramid (LP), Discrete Wavelet Transform (DWT) and Dual-Tree Complex Wavelet Transform (DT-CWT) (please see [6] for more details on these techniques). In the multiresolution methods (LP, DWT and DT-CWT) a 5-level decomposition is used and fusion is performed by selecting the coefficient with the maximum absolute value, except for the case of the lowest resolution subband where the mean value is used. For all sequences apart from the grey-scale fusion results, colour fusion results were computed in the YUV domain.

3.1. Video Fusion Assessment

Several objective performance measures for image fusion have also been proposed where the knowledge of ground-truth is not assumed. The measure used as the basis for the Piella metric and the Bristol metric is SSIM.

3.1.1 Piella Metric

Since images are generally non-stationary signals, it is appropriate to measure SSIM Q_0 over local regions and then combine the different results into a single measure Q. For each window w the local quality index $Q_0(X, Y|w)$ is computed for the pixels within the sliding window w.

$$Q(X,Y) = \frac{1}{|W|}\sum_{w\in W} Q_0(X,Y|w) \tag{10}$$

where W is the family of all windows and $|W|$ is the cardinality of W. In order to apply the SSIM to image fusion evaluation, Piella and Heijmans [14] introduce salient information to the metric:

$$Q_p(X,Y,F) = \sum_{w\in W} c(w)[\lambda Q(X,F|w)+(1-\lambda)Q(Y,F|w)] \tag{11}$$

where X and Y are the input images, F is the fused image, $c(w)$ is the overall saliency of a window and λ is defined as:

$$\lambda = \frac{s(X|w)}{s(X|w)+s(Y|w)} \tag{12}$$

and should reflect the relative importance of image X compared to image Y within the window w. Finally, to take into account aspects of the human visual system (HVS), the same measure is computed with "edge images" (X', Y' and F') instead of the grey-scale images X, Y and F.

$$Q_E(X,Y,F) = Q_p(X,Y,F)^{1-\alpha} Q_p(X',Y',F')^{\alpha} \tag{13}$$

3.1.2 Petrovic Metric

The fusion metric proposed by Petrovic and Xydeas [13], is obtained by evaluating the relative amount of edge information transferred from the input images to the output image. It uses a Sobel edge operator to calculate the strength and orientation information of each pixel in the input and output images. The relative strength and orientation "change" values, $G_{XF}(n,m)$ and $A_{XF}(n,m)$ respectively, of an input image X with respect to the fused one F are calculated. These measures are then used to estimate the edge strength and orientation preservation values, $Q_g^{XF}(n,m)$ and $Q_\alpha^{XF}(n,m)$ respectively. The overall edge information preservation values are then defined as:

$$Q^{XF}(n,m) = Q_g^{XF}(n,m)\cdot Q_\alpha^{XF}(n,m) \tag{14}$$

where $0 \le Q^{XF}(n,m) \le 1$. Having calculated $Q^{XF}(n,m)$ and $Q^{YF}(n,m)$, a normalised weighted performance metric of a given process p that fuses X and Y into F is given by:

$$Q_p = \frac{\sum_{n=1}^{N}\sum_{m=1}^{M} Q^{XF}(n,m)w_X(n,m) + Q^{YF}(n,m)w_Y(n,m)}{\sum_{n=1}^{N}\sum_{m=1}^{M} w_X(n,m) + w_Y(n,m)} \tag{15}$$

The edge preservation values $Q^{XF}(n,m)$ and $Q^{YF}(n,m)$ are weighted by coefficients $w_X(n,m)$ and $w_Y(n,m)$, which reflect the perceptual importance of the corresponding edge elements within the input images. Note that in this method, the visual information is associated with the edge information while the region information is ignored.

3.1.3 Bristol Metric

In the computation of Piella metric, the parameter λ in equation 12 is computed with $s(X|w)$ and $s(Y|w)$ being the variance (or the average in the edge images) of images X and Y within window w, respectively. Therefore, there is no clear measure of how similar each input image is to the final fused image. A novel fusion performance measure was proposed in [4] that takes into account the similarity between the input image block and the fused image block within the same spatial position. It is defined as:

$$Q_b = \sum_{w\in W} sim(w)Q(X,F|w) + (1-sim(w))Q(Y,F|w) \tag{16}$$

where X and Y are the input images, F is the fused image, w is the analysis window and W is the family of all windows. $sim(X,Y,F|w)$ is defined as:

$$sim(w) = \begin{cases} 0 & \text{if } \frac{\sigma_{xf}}{\sigma_{xf}+\sigma_{yz}} < 0 \\ \frac{\sigma_{xf}}{\sigma_{xf}+\sigma_{yz}} & \text{if } 0 \le \frac{\sigma_{xf}}{\sigma_{xf}+\sigma_{yz}} \le 1 \\ 1 & \text{if } \frac{\sigma_{xf}}{\sigma_{xf}+\sigma_{yz}} > 1 \end{cases} \tag{17}$$

Each analysis window is weighted by the $sim(w)$ that is dependent on the similarity in the spatial domain between the input image and the fused image. The $sim(w)$ function is designed to have an upper limit of one, so that the impact of less significant blocks is completely eliminated when the other input blocks' similarity measure equals one.

4. Experimental Results

4.1. Dataset Description

Four sequences were used in this study: *Eden 2.1*; *Eden 4.1*; *QQ*; and *OTCBVS*. The selected multimodal video sequences have different scene complexity, different illumination levels, various target/object sizes and distances from target/object to sensors. In addition, three of the selected

video sequences are real-life videos and one (*QQ*) is a synthetic video sequence. The sequences are now briefly described from an object tracking perspective:

Eden 2.1 : Camouflage man walking through an opening, against a leafy background until obscured by a tree. Available at www.imagefusion.org, more details in [8]

Eden 4.1 : Man in white T-shirt walking left-right and then right-left. Partially obscured on a number of occasions by other people. Available at www.imagefusion.org, more details in [8]

QQ : Airborne view. Target is a small vehicle (tractor) towards which the sensor is heading. This data set was generated and kindly provided to us by QinetiQ, UK..

OTCBVS : Man walking through a university campus. Walks through shadow, also variable cloud cover. Some occlusion by other people. Available at http://www.cse.ohio-state.edu/OTCBVS-BENCH, more details in [5].

4.2. Tracking Results

Four types of image sequences were used: Visible; Infrared; fusion of visible and IR using averaging (AVE); and fusion of visible and IR using the complex wavelet transform (DT-CWT). All four types were used for all data sets, i.e. *Eden 2.1*, *Eden 4.1*, *OTCBVS* and *QQ* sequences. The first frames for each sequence are shown in Figures 1, 2, 3 and 4.

Figure 1. Initial frame from *Eden 2.1* sequence. Modalities are (clockwise from top left): VIZ; IR; DT-CWT and AVE

We now consider qualitatively how the tracking algorithm performs for the different sequences. Due to space constraints, it is not possible to provide further illustrations for all the sequences; it is hoped a textual description will suffice.

Figure 2. Initial frame from *Eden 4.1* sequence. Modalities are (clockwise from top left): VIZ; IR; DT-CWT and AVE

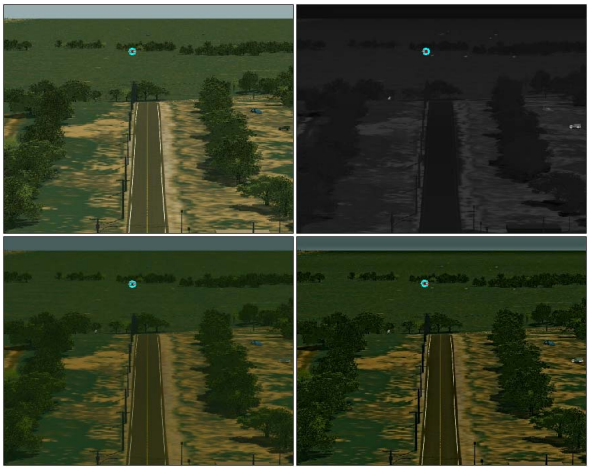

Figure 3. Initial frame from *QQ* sequence. Modalities are (clockwise from top left): VIZ; IR; DT-CWT and AVE

Eden 2.1 : In this sequence, the target is very similar to the background in the visible spectrum, and relatively dissimilar in the IR. For this reason, the target is quickly lost in the visible sequence. All the other modalities perform much better. However, the IR performs best due to the targets' significant separation from the background. The fused sequences effectively reduce this separation by the inclusion of visible information.

Eden 4.1 : Due to the occlusion by other targets with similar IR characteristics, it was found that often the visible mode is the most discriminating. For this reason, the visible sequence gives the most accurate tracking, with the fusion methods next, followed by IR.

QQ : The initial small size of the target means that in the visible spectrum, the appearance changes signif-

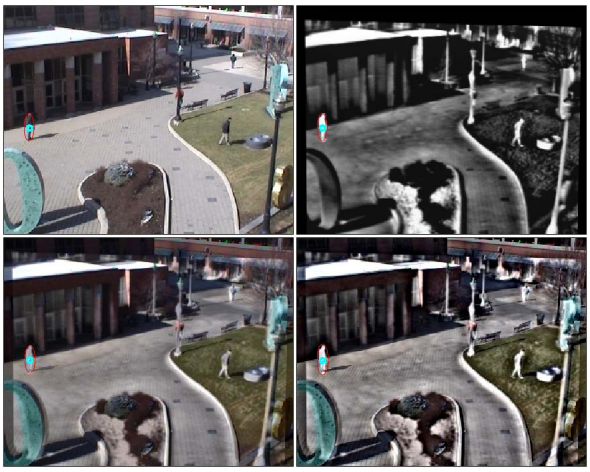

Figure 4. Initial frame from *OTCBVS* sequence. Modalities are (clockwise from top left): VIZ; IR; DT-CWT and AVE

icantly as the sensor approaches the target. This ultimately causes the visible tracker to fail. In the IR spectrum, the appearance is much more constant and the IR tracker works well. Note that we do not currently update the target model. The combination of the two modalities for the AVE tracker predictably leads to a tracking performance somewhere between IR and VIZ. The higher contrast of the DT-CWT algorithm leads to a greater change in the target appearance, thus lock is lost relatively quickly.

OTCBVS : The visible tracker is quickly confused by the changes in lighting, something which does not affect the IR and fused sequences. Averaged over a number of noise realisations, there is relatively little to choose between the IR and fused sequences; all are capable of tracking through to the end of the sequence.

Tracking results were also computed for monochrome visible spectrum images. The motivation behind this is that this is what is used to calculate the metrics presented in Section 4.3. In general, this degrades the quality of the tracking; even if lock is maintained, the accuracy with which the target is tracked is often reduced. However, in some circumstances it can lead to an improvement in the tracking. This is typically in scenes where the change in the chrominance of the target is greater than the luminance e.g. the *QQ* sequence, and to a lesser extent *OTCBVS*. The drawback to this approach is that the effects of clutter may be increased.

In summary, the results obtained in this section suggest strongly that the choice of sensor modality (or fusion thereof) should reflect the types of target and clutter that are expected. On average, the IR mode was found to be the most useful. However, when the target is in the proximity of e.g. other people, it can fail quickly, particularly in situations where the targets are sufficiently small such that the different thermal properties of e.g. arms, torso, head cannot be resolved. Under these circumstances, fusion, particularly the AVE approach, is beneficial. In addition, in a situation when the task is not to simply track a single target, but to determine/estimate its position with respect to another object that is not visible in the IR video, video fusion is essential in order to perform the task successfully and accurately.

4.3. Videos Fusion Assessment Using Standard Image Fusion Metrics

The first step in the evaluation of pixel-level fusion methods was to calculate standard image fusion metrics [14, 13, 4] on a frame-by-frame basis using the grey-scale versions of the visible videos and grey-scale video fusion. For each frame from the fused video and the two input videos (VIZ and IR) we calculate a fusion metric value. Tables 1-4 give an overview of the results, presenting the mean values of the standard fusion metrics over the full length of the fused videos. It is obvious that there is no significant difference between the performance of the fusion methods on the real-life videos (*Eden 2.1*, *Eden 4.1* and *OTCBVS*) and the synthetic video data (*QQ*). For the real-life video data, the overall best performing fusion method is DT-CWT, followed by either DWT or LP, while the averaging method scores significantly lower.

Table 1. Video fusion performance, *Eden 2.1* sequence, the mean value of the metric over 107 frames

Fusion	Fusion method			
Metric	AVE	LP	DWT	DT-CWT
Piella	0.835	0.817	0.882	**0.888**
Petrovic	0.268	0.368	0.415	**0.432**
Bristol	0.616	0.702	0.697	**0.713**

Table 2. Video fusion performance *Eden 4.1* sequence, the mean value of the metric over 193 frames

Fusion	Fusion method			
Metric	AVE	LP	DWT	DT-CWT
Piella	0.831	**0.906**	0.882	0.897
Petrovic	0.324	0.652	0.562	**0.654**
Bristol	0.599	0.744	0.722	**0.748**

Visual comparison between methods confirms the conclusion derived using fusion metrics—the multiresolution methods are superior to AVE in terms of the amount of visual information transferred from the input videos to the fused video. For example, in Fig. 1 it is clear that the vegetation detail from the visual image is far better transferred into the fused image by the multiresolution method than in

AVE. In addition, the texture of the images fused using multiresolution methods is visually more pleasing and the overall contrast of the fused images is much better if fusion is performed using LP, DWT and especially DT-CWT.

In addition, when synthetic video inputs are fused, all the fusion assessment metrics rank DT-CWT as the best performing fusion method. Therefore, even in the case of the synthetic video, metrics give consistent results despite the synthetic texture of the objects and the "unnatural" contrast in the input videos.

Table 3. Video fusion performance *OTCBVS* sequence, the mean value of the metric over 375 frames

Fusion Metric	Fusion method			
	AVE	LP	DWT	DT-CWT
Piella	0.883	0.834	0.895	**0.901**
Petrovic	0.421	0.402	0.534	**0.563**
Bristol	0.682	0.598	0.664	**0.696**

Table 4. Video fusion performance *QQ* sequence, the mean value of the metric over 162 frames

Fusion Metric	Fusion method			
	AVE	LP	DWT	DT-CWT
Piella	0.837	0.800	0.845	**0.846**
Petrovic	0.567	0.612	0.761	**0.779**
Bristol	0.725	0.667	0.732	**0.746**

4.4. Correlation Between Tracking Results and Qualitative and Quantitative Fused Video Assessment Results

The presented experimental results clearly demonstrate that there is a low correlation between the objective assessment and the tracking results. Namely, the best tracking performance is generally attained if either simple AVE fusion method is implemented or if fusion is not performed at all and only the IR input is used. On the other hand, metrics for fusion assessment clearly point towards the supremacy of the multiresolution methods, especially DT-CWT. It seems that the fusion assessment metrics correspond well with the subjective quality of the fused videos, as it is obvious that the multiresolution methods produce fused videos with better contrast, more visible details and all the salient features transferred from input to fused videos. Conversely, experiments have shown that a better visual quality of a fused video does not guarantee a better tracking performance, as often the simple AVE method outperforms the DT-CWT method. Therefore, a new, tracking-oriented, video fusion metric is needed that is better able to model the tracking performance for a fused video sequence.

5. Conclusions

This paper presents an experimental approach to the assessment of the effects of pixel-level fusion on object tracking in multimodal surveillance videos. The results obtained in the experiments strongly suggest that on average, the IR mode is the most useful when it comes to tracking objects that are well seen in the IR spectrum (e.g. humans). However, under some circumstances, video fusion, typically the AVE approach, is beneficial. The underlying cause of these observations is linked to the distance between the foreground and background pixel distributions. Fusing two modalities together cannot be guaranteed to increase this distance; indeed, fusion may well decrease it.

In contrast, in a situation where the task is to track multiple targets, and the targets are most separable from the background (and possibly each other) in different modalities, then video fusion is essential in order to perform the task successfully and accurately. This is due to the inclusion of complementary and contextual information from all input sources, making it more suitable for further analysis by either a human observer or a computer program. Current metrics for fusion assessment are geared towards the human observer, and clearly point towards the supremacy of the multiresolutional methods, especially DT-CWT. As this does not correlated with the results obtained, a new tracking-oriented metric is needed that would be able to reliably assess the tracking performance for a fused video sequence.

6. Acknowledgements

This work has been funded by the UK Data and Information Fusion Defence Technology Centre (DIF DTC) AMDF and Tracking cluster projects. We would like to thank the Eden Project for allowing us to record the Eden Project Multi-Sensor Data Set (of which Eden 2.1 and 4.1 are part of) and QinetiQ, UK, for providing the *QQ* data set.

References

[1] M. Arulampalam, S. Maskell, N. Gordon, and T. Clapp. A tutorial on particle filters for on-line nonlinear/non-Gaussian Bayesian tracking. *IEEE Transactions on Signal Processing*, 50:174–188, 2002.

[2] P. Brasnett, L. Mihaylova, N. Canagarajah, and D. Bull. Sequential Monte Carlo tracking by fusing multiple cues in video sequences. *Image and Vision Computing*, (in print).

[3] D. Comaniciu, V. Ramesh, and P. Meer. Kernel-based object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25:564–577, 2003.

[4] N. Cvejic, A. Loza, D. Bull, and N. Canagarajah. A similarity metric for assessment of image fusion. *International Journal of Signal Processing*, 2:178–182, 2005.

[5] J. Davis and V. Sharma. Fusion-based background-subtraction using contour saliency, 2005. Proc. IEEE International Workshop on Object Tracking and Classification Beyond the Visible Spectrum.

[6] T. Dixon, J. Li, J. Noyes, T. Troscianko, S. Nikolov, J. Lewis, E. Canga, D. Bull, and N. Canagarajah. Scanpath analysis of fused multi-sensor images with luminance change, 2006. Proc. International Conference on Information Fusion.

[7] M. Isard and A. Blake. Condensation-conditional density propagation for visual tracking. *International Journal of Computer Vision*, 29:5–28, 1998.

[8] J. Lewis, S. Nikolov, A. Loza, E.-F. Canga, N. Cvejic, J. Li, A. Cardinali, N. Canagarajah, D. Bull, T. Riley, D. Hickman, and M. I. Smith. The Eden Project multi-sensor data set, 2006. Technical Report TR-UoB-WS-Eden-Project-Data-Set, University of Bristol, Waterfall Solutions Ltd.

[9] A. Loza, L. Mihaylova, N. Canagarajah, and D. Bull. Structural similarity measure for object tracking in video sequences, 2006. Proc. International Conference on Information Fusion.

[10] L. Mihaylova, A. Loza, S. Nikolov, J. Lewis, E. Canga, J. Li, D. Bull, and N. Canagarajah. The influence of multi-sensor video fusion on object tracking using a particle filter, 2006. Proc. Workshop on Multiple Sensor Data Fusion.

[11] K. Nummiaro, E. B. Koller-Meier, and L. V. Gool. An adaptive color-based particle filter. *Image and Vision Computing*, 21:99–110, 2003.

[12] P. Pérez, C. Hue, J. Vermaak, and M. Gangnet. Color-based probabilistic tracking, 2002. Proc. European Conference on Computer Vision.

[13] V. Petrovic and C. Xydeas. Objective evaluation of signal-level image fusion performance. *Optical Engineering*, 44, 2005.

[14] G. Piella and H. Heijmans. A new quality metric for image fusion, 2003. Proc. IEEE International Conference on Image Processing.

[15] C. Shen, A. van den Hengel, and A. Dick. Probabilistic multiple cue integration for particle filter based tracking, 2003. Proc. Digital Image Computing: Techniques and Applications.

[16] A. Toet and E. Franken. Fusion of visible and thermal imagery improves situational awareness. *Displays*, 18:85–95, 1997.

[17] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 2004.